German Human Genome Archive (GHGA)

Sprecher/in: Oliver Stegle, DKFZ, o.stegle@dkfz-heidelberg.de
Board of Directors: Oliver Kohlbacher, Univ. Tübingen, oliver.kohlbacher@uni-tuebingen.de, Jan Korbel, EMBL, Jan Korbel jan.korbel@embl.org, Eva Winkler, Univ. Klinikum Heidelberg, eva.winkler@med.uni-heidelberg.de

Forschungsgebiet Menschliche Genomdaten und andere verwandte Omics-Daten sind integraler Bestandteil der biomedizinischen Forschung und die Versorgung von morgen. Der Bedarf Omics-Daten offen und FAIR für die Forschung nutzen zu können wird balanciert durch die Notwendigkeit, diese Daten sicher und geschützt aufzubewahren und nur zu legitimen Forschungszwecken zugreifbar zu machen. Existierende Infrastrukturen, insbesondere das Europäische Genom-Phenom-Archiv (EGA) können die spezifischen Anforderungen des Deutschen Rechts nur ungenügend abbilden. Darüber hinaus ist EGA als Archiv konzipiert und bedient insbesondere nicht die Wünsche der Forschergemeinde nach effizienten, benutzerfreundlichen Analysen im großen Maßstab und zur Replikation von Ergebnissen über verschiedene Kohorten hinweg. Das Fehlen einer technisch und rechtlich sicheren nationalen Infrastruktur für Omics-Daten ist ein wesentliches Hindernis des Potentials von existierenden und zukünftigen Genomdateninitiativen in Deutschland und Europa voll auszuschöpfen.

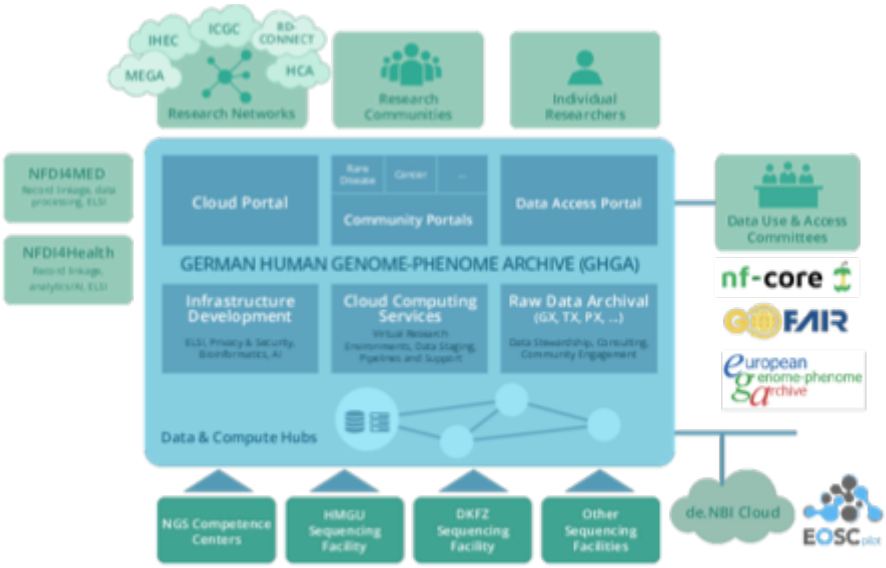

Abbildung 1: Übersicht über die Struktur des Konsortiums und die Einbettung in nationale und internationale Initiativen.

Ziele Das Deutsche Humangenom-Phenom-Archiv (GHGA) wird diesen Bedarf bedienen und eine nationale Infrastruktur für die sichere Speicherung, Zugriff und Analyse menschlicher Omics-Daten (z.B. Genome, Transkriptome) in einem einheitlichen ethisch-rechtlichen Rahmen aufbauen. GHGA setzt dabei auf existierende nationale Omics-Datenlieferanten und deren IT-Infrastrukturen, um eine harmonisierte, interoperable Infrastruktur zu schaffen. Die direkte Verbindung zu großen Omics-Zentren wird die Hürde zur Archivierung der Daten reduzieren, indem direkter Metadatentransfer zu GHGA ermöglicht wird. GHGA wird als nationaler Knoten in die zukünftige föderierte

EGA-Infrastruktur eingebunden. Dadurch werden Forscher in Deutschland in die Lage versetzt, internationale Standards zum Datenaustausch stärker mitzugestalten und führende Rollen in internationalen Forschungsnetzwerken einzunehmen (z.B. *1+ Million European Genomes Initiative* [MEGA]).

Anforderungen an das Forschungsdatenmanagement Kernanforderung an GHGA als Omics-Forschungsdateninfrastruktur ist die Funktion als ein Archiv für Genomdaten. Darüber hinaus wird GHGA Funktionalität jenseits reiner Archivierung anbieten, um insbesondere den Zugriff auf populationsweite Omics-Datensätze zu demokratisieren. Sichere, private Cloud-Infrastrukturen sollen rechenintensive Anwendungen (inklusive KI-Anwendungen) auf diesen Datenbeständen ermöglichen, ohne die Notwendigkeit eines Daten-Downloads. Maßgeschneiderte Portale für bestimmte Nutzergruppen sowie die Kuratierung von besonders interessanten Referenzdatensätzen werden den Wert der Infrastruktur für Forscher und Ärzte gleichermaßen erhöhen - diese Nutzer werden umgekehrt auch die Entwicklung der Infrastruktur maßgeblich mitgestalten.
Die initiale Schwerpunktsetzung auf Krebs und seltene Erkrankungen ist in Übereinstimmung mit den Schwerpunkten anderer Initiativen, insbesondere der geplanten Deutschen Genominitiative (genomDE). Aus-/Weiterbildungsaktivitäten werden den Nachwuchs früh an GHGA und Omics-Datenanalyse heranführen.

Geplante Maßnahmen und Services GHGA adressiert die genannten Anforderungen durch ein integriertes Konzept, dass Maßnahmen mit konzeptionellem Fokus, Communities, Daten, sowie dem Betrieb einer Datenplattform eng verknüpft (Abb. 2).

Abbildung 2: Übersicht über die in GHGA geplanten Maßnahmen.

Kernziel von GHGA ist der Betrieb einer verteilten Dateninfrastruktur (TA C & TA D; Abb. 2), die Omics-Daten die national generiert innerhalb des föderierten EGA-Modells zur Verfügung stellt. In diesem Kontext soll insbesondere auch eine enge Anbindung an akademische Cloud-Dienste etabliert werden, um damit eine sichere und kontrollierte Verarbeitung von Omics Daten durch eine breite Gruppe an Nutzern zu ermöglichen. TA C wird darüber hinaus Community Standards zur Prozessierung dieser Daten erarbeiten, sowie interaktive Portale für den einfachen und sicheren Zugriff auf Omics Daten etablieren. Weitere Maßnahmen sind die Interaktion mit und das Entwickeln von Netzwerken für die Interaktion mit den

Communities (TA A), um einen starken Anwendungsbezug der Dienste sicherzustellen. Konzeptionelle Themen, wie die Entwicklung eines ethisch-rechtlichen Rahmens sowie die Implementation von FAIR Prinzipien werden in TA B adressiert.

Datenarten GHGA hat einen definierten Fokus auf humane Omics-Daten, die aufgrund ihres Personenbezugs besonders strikten Datenschutzanforderungen unterliegen. Neben genomischen Sequenzdaten bezieht dies weitere durch Sequenzierung gewonnen Datentypen mit ein, wie beispielsweise RNA-Daten, epigenetische Daten aber auch Proteomics. GHGA wird sowohl konventionelle Omics-Daten verarbeiten, als auch die durch moderne Methoden zur Einzelzellanalytik erzeugte Daten.

Rolle innerhalb der NFDI und Interaktion mit anderen Konsortien GHGA hat als Basisinfrastruktur für Omics Daten eine definierte Rolle innerhalb der NFDI und wird mit zahlreichen spezialisierten Konsortien zusammenarbeiten. Bisherige geplante Interaktionen zu Omics-Daten sind mit den folgenden Konsortien vorgesehen: NFDI4Health, NFDI4Med & NFDI4Microbiota. Darüber hinaus wird GHGA Datendienste zu Omics-Daten grundsätzlich allen NFDI-Konsortien zur Verfügung stellen.

Querschnittsthemen
Alle wesentlichen Themenblöcke in GHGA werden von weitreichenden Maßnahmen zur Zusammenarbeit profitieren. Wir sehen besonderem Bedarf zu einer weitreichenden Integration bei ethisch-rechtlichen Fragestellungen aber auch hinsichtlich Strukturen um Kapazitäten für das wissenschaftliche Rechnen und die langfristige Datenspeicherung zu etablieren und gemeinschaftlich über NFDI-Konsortien hinweg zu nutzen. Spezifische Querschnittsthemen, die von GHGA bearbeitete werden und daher von besonderem Interesse sind:

- Standardisierung von Phänotypen und sichere Verknüpfung mit medizinischen Daten.
- Management von Einwilligungen und ethisch-rechtliche Grundlagen für die Datenverarbeitung.
- Cloud Computing und Datenplattformen für das wissenschaftlichen Rechnen.
- Standardisierung von Daten Prozessieren und Workflow Management.

Erwartungen an die NFDI Konferenz
GHGA ist an einem Austausch zu Scherschnitt Fragen interessiert, sowie neue Verknüpfungen mit weiteren in der Entwicklung befindlichen Konsortien.

Vorgesehene Mitglieder des Konsortiums (Co-Sprecherinnen/Co-Sprecher und die weiteren, beteiligten Institutionen):

Co-Sprecher/in	Zugehörige Institution
Peer Bork Co-lead Maßnahme B3 bork@embl.org	Europäisches Laboratorium für Molekularbiologie, Heidelberg
Ivo Buchhalter Co-lead Maßnahme C1,C5 i.buchhalter@dkfz-heidelberg.de	Deutsches Krebsforschungszentrum, Heidelberg
Andreas Dahl Co-lead Maßnahme D2 andreas.dahl@tu-dresden.de	Technische Universität Dresden
Julien Gagneur Co-lead Maßnahme C2 gagneur@in.tum.de	Technische Universität München
Wolfgang Huber Co-lead Maßnahme A3 whuber@embl.de	Europäisches Laboratorium für Molekularbiologie, Heidelberg
Daniel Hübschmann Co-lead Maßnahme C3 d.huebschmann@dkfz.de	Deutsches Krebsforschungszentrum, Heidelberg
Oliver Kohlbacher Co-lead Maßnahme A3, C3, D1, E1 oliver.kohlbacher@uni-tuebingen.de	Eberhard-Karls-Universität, Tübingen
Jan Korbel Co-lead Maßnahme C4, E2 korbel@embl.de	Europäisches Laboratorium für Molekularbiologie, Heidelberg
Martin Lablans Co-lead Maßnahme B2 m.lablans@dkfz.de	Deutsches Krebsforschungszentrum, Heidelberg
Ulrich Lang Co-lead Maßnahme C5, D1, D2 lang@uni-koeln.de	Universität zu Köln
Peter Lichter Co-lead Maßnahme A1 peter.lichter@dkfz-heidelberg.de	Deutsches Krebsforschungszentrum, Heidelberg
Fruzsina Molnár-Gábor Co-lead Maßnahme B1 fruzsina.molnar-gabor@adw.uni-heidelberg.de	Heidelberger Akademie der Wissenschaften, Heidelberg
Susanne Motameny Co-lead Maßnahme C1 susanne.motameny@uni-koeln.de	Universität zu Köln
Sven Nahnsen Co-lead Maßnahme B2 sven.nahnsen@qbic.uni-tuebingen.de	Eberhard-Karls-Universität, Tübingen

Weitere beteiligte Institutionen:

Mitarbeiter	Zugehörige Institution
Thomas Keane tk2@ebi.uk.ac	European Bioinformatics Institute (EBI), Cambridge, Vereinigtes Königreich
Mario Fritz, Ninja Marnau Fritz@cispa.saarland, marnau@cispa.saarland	Helmholtz Zentrum für Informationssicherheit, Saarbrücken
Stephan Hachinger stephan.hachinger@lrz.de	Leibniz Rechenzentrum d. Bayerischen Akademie d. Wissenschaftern, München
Alice HcHardy alice.mchardy@helmholtz-hzi.de	Helmholtz Zentrum für Infektionsforschung, Braunscherig
Stefan Fröhling und Hanno Glimm stefan.froehling@nct-heidelberg.de, hanno.glimm@nct-dresden.de	Nationales Centrum für Tumorerkrankungen, Heidelberg und Dresden